Sections 7.1, 7.2, 7.4, & 7.6

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Chapter 7 example: Body fat

$n = 20$ healthy females 25–34 years old.

- x_1 = triceps skinfold thickness (mm)
- x_2 = thigh circumference (cm)
- x_3 = midarm circumference (cm)
- Y = body fat (%)

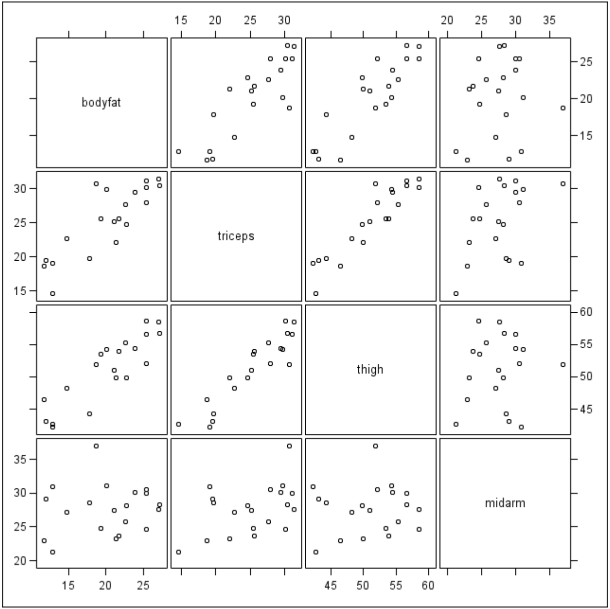
Obtaining Y_i, the percent of the body that is purly fat, requires immersing a person in water. Want to develop model based on simple body measurements that avoids people getting wet.

SAS code

```
*******************************
*  Body fat data from Chapter 7
*******************************;
data body;
  input triceps thigh midarm bodyfat @@;
  cards;
  19.5  43.1  29.1  11.9  24.7  49.8  28.2  22.8
  30.7  51.9  37.0  18.7  29.8  54.3  31.1  20.1
  19.1  42.2  30.9  12.9  25.6  53.9  23.7  21.7
  31.4  58.5  27.6  27.1  27.9  52.1  30.6  25.4
  22.1  49.9  23.2  21.3  25.5  53.5  24.8  19.3
  31.1  56.6  30.0  25.4  30.4  56.7  28.3  27.2
  18.7  46.5  23.0  11.7  19.7  44.2  28.6  17.8
  14.6  42.7  21.3  12.8  29.5  54.4  30.1  23.9
  27.7  55.3  25.7  22.6  30.2  58.6  24.6  25.4
  22.7  48.2  27.1  14.8  25.2  51.0  27.5  21.1
;

proc sgscatter; matrix bodyfat triceps thigh midarm; run;
```

Scatterplot



Correlation coefficients

```
proc corr data=body; var triceps thigh midarm; run;
```

```
    Pearson Correlation Coefficients, N = 20
           Prob > |r| under H0: Rho=0

              triceps         thigh        midarm

triceps       1.00000       0.92384       0.45778
                             <.0001        0.0424

thigh         0.92384       1.00000       0.08467
               <.0001                      0.7227

midarm        0.45778       0.08467       1.00000
               0.0424        0.7227
```

There is high correlation among the predictors. For example $r = 0.92$ for triceps and thigh. These two variables are *essentially carrying the same information*. Maybe only one or the other is really needed.

In general, one predictor may be essentially perfectly predicted by the remaining predictors (a high "partial correlation"), and so would be unecessary if the other predictors are in the model.

7.1 Extra sums of squares

"Extra" sums of squares are defined as the difference in SSE between a model with some predictors and a larger model that adds *additional* predictors.

Fact: As predictors are added, the SSE can only decrease. The extra sums of squares is how much the SSE decreases:

def'n Let $x_1, x_2, \ldots, x_k$ be predictors in a model.

$$SSR(x_{j+1}, \ldots, x_k | x_1, x_2, \ldots, x_j) = SSE(x_1, x_2, \ldots, x_j) - SSE(x_1, x_2, \ldots, x_j, x_{j+1}, \ldots, x_k),$$

the difference in the sums of squared errors from the reduced to the full model.

This is how much of the total variation in SSTO is further explained by adding the new predictors.

Example with $k = 8$ predictors

The predictors under consideration are

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8.$$

There are two models

$$\begin{aligned} \text{Reduced :} &\quad x_1, x_3, x_5, x_6, x_8 \\ \text{Full :} &\quad x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \end{aligned}$$

$$\begin{aligned} \text{Extra SS} &= SSR(x_2, x_4, x_7 | x_1, x_3, x_5, x_6, x_8) \\ &= SSE(\text{reduced}) - SSE(\text{full}) \\ &= SSE(x_1, x_3, x_5, x_6, x_8) - SSE(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) \\ &= SSR(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) - SSR(x_1, x_3, x_5, x_6, x_8) \end{aligned}$$

This is how much *additional* total variability (SSTO) is explained by adding x_2, x_4, x_7 to a model that already has x_1, x_3, x_5, x_6, x_8.

7.2 Associated tests

We can formally test whether a certain set of predictors is useless, *in the presence* of other predictors in the model. This is the *general linear test* we talked about a few lectures ago (in simple linear regression).

In the example above, we can test whether x_2, x_4, x_7 are needed if x_1, x_3, x_5, x_6, x_8 are in the model. If full (with $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$) model has much lower SSE than reduced model (without x_2, x_4, x_7) then at least one of x_2, x_4, x_7 is needed.

F-test

Say we want to test whether we can drop q variables from a model that has $p = k + 1$ (including the intercept), $q < p$.

Let R denote the reduced model and F the full, and $SSE(R)$, $SSE(F)$ denote the sums of squared errors from the two models. To test $H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_q} = 0$ in the full model

$$\begin{aligned} F^* &= \frac{[SSE(R) - SSE(F)]/q}{SSE(F)/(n-p)} \\ &\sim F(q, n-p) \end{aligned}$$

If $H_0 : \beta_{j_1} = \beta_{j_2} = \cdots = \beta_{j_q} = 0$ is true; a p-value for the test is $P(F(q, n-p) > F^*)$.

Can carry this out in SAS using `test` in `proc reg`.

F-test example with $k = 8$ predictors

To test $H_0 : \beta_2 = \beta_4 = \beta_7 = 0$,

$$\begin{aligned}
F^* &= \frac{[SSE(\text{reduced}) - SSE(\text{full})]/(\#\text{ parameters in test})}{MSE(\text{full})} \\
&= \frac{[SSE(x_1, x_3, x_5, x_6, x_8) - SSE(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)]/3}{SSE(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)/(n-9)} \\
&= \frac{SSR(x_2, x_4, x_7 | x_1, x_3, x_5, x_6, x_8)/3}{SSE(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)/(n-9)} \\
&\sim F(3, n-9)
\end{aligned}$$

if $H_0 : \beta_2 = \beta_4 = \beta_7 = 0$ is true.

Bodyfat example

```
proc reg data=body;
 model bodyfat=triceps thigh midarm;
 test thigh=0, midarm=0; run;
```

```
       Test 1 Results for Dependent Variable bodyfat

                                Mean
Source             DF         Square    F Value    Pr > F
Numerator           2       22.35741       3.64    0.0500
Denominator        16        6.15031
```

Reject $H_0 : \beta_2 = \beta_3 = 0$ in

$$\text{fat}_i = \beta_0 + \beta_1 \text{triceps}_i + \beta_2 \text{thigh}_i + \beta_3 \text{midarm}_i + \epsilon_i$$

with $p = 0.05$.

Type I (sequential) sums of squares

Note (pp. 260–262): Say you have $k = 4$ predictors. Then the SSR for the full model can be written

$$\begin{aligned} SSR &= SSR(x_1, x_2, x_3, x_4) \\ &= SSR(x_1) + SSR(x_2|x_1) + SSR(x_3|x_1, x_2) + SSR(x_4|x_1, x_2, x_3). \end{aligned}$$

These are called *sequential sums of squares*, or Type I sums of squares. They explain how much variability is soaked up by adding

predictors sequentially to a model. There are four corresponding hypothesis tests with these sequential sums of squares:

Model	Hypothesis	F-statistic
$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$	$H_0 : \beta_1 = 0$	$\frac{SSR(x_1)}{MSE(x_1)}$
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$	$H_0 : \beta_2 = 0$	$\frac{SSR(x_2 \vert x_1)}{MSE(x_1,x_2)}$
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$	$H_0 : \beta_3 = 0$	$\frac{SSR(x_3 \vert x_1,x_2)}{MSE(x_1,x_2,x_3)}$
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$	$H_0 : \beta_4 = 0$	$\frac{SSR(x_4 \vert x_1,x_2,x_3)}{MSE(x_1,x_2,x_3,x_4)}$

Bodyfat example

You can get sequential SS from `proc reg` by adding `ss1` as a `model` option. `proc glm` gives them automatically.

```
proc glm data=body;
 model bodyfat=triceps thigh midarm / solution; run;
-----------------------------------------------------------------------------------
Source                     DF       Type I SS     Mean Square    F Value    Pr > F
triceps                     1     352.2697968     352.2697968      57.28    <.0001
thigh                       1      33.1689128      33.1689128       5.39    0.0337
midarm                      1      11.5459022      11.5459022       1.88    0.1896
-----------------------------------------------------------------------------------
```

- Reject $H_0 : \beta_1 = 0$ in $\text{fat}_i = \beta_0 + \beta_1 \text{triceps}_i + \epsilon_i$ with $p < 0.0001$.
- Reject $H_0 : \beta_2 = 0$ in $\text{fat}_i = \beta_0 + \beta_1 \text{triceps}_i + \beta_2 \text{thigh}_i + \epsilon_i$ with $p = 0.034$.
- Accept $H_0 : \beta_3 = 0$ in $\text{fat}_i = \beta_0 + \beta_1 \text{triceps}_i + \beta_2 \text{thigh}_i + \beta_3 \text{midarm}_i + \epsilon_i$ with $p = 0.190$.
- Order entered (triceps, thigh, midarm) matters!

ANOVA table & decomposing the SSR(F)

```
                                        Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        3     396.9846118     132.3282039      21.52    <.0001
Error                       16      98.4048882       6.1503055
Corrected Total             19     495.3895000
```

The sequential extra sums of squares is given on the last slide: $SSR(x_1) = 352.3$; $SSR(x_2|x_1) = 33.2$, and $SSR(x_3|x_1, x_2) = 11.5$. Almost all of the $SSR(x_1, x_2, x_3) = 397.0$ is explained by x_1 (triceps) alone.

Also note, as required,

$$SSR(x_1, x_2, x_3) = 397.0 = 352.3+33.2+11.5 = SSR(x_1)+SSR(x_2|x_1)+SSR(x_3|x_1, x_2).$$

Finally, we strongly reject $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

7.4 Coefficients of partial determination

We can *standardize* extra sums of squares to be between 0 and 1 (like R^2).

The **coefficient of partial determination** is the fraction by which the sum of squared errors is reduced by adding predictor(s) to an existing model. Examples:

- $R^2_{Y2|1} = SSR(x_2|x_1)/SSE(x_1)$
- $R^2_{Y3|12} = SSR(x_3|x_1, x_2)/SSE(x_1, x_2)$
- $R^2_{Y32|1} = SSR(x_2, x_3|x_1)/SSE(x_1)$

For example, if $R^2_{Y3|12} = 0.5$ then *50% of the remaining variability* is explained by adding x_3 to a model that already had x_1 and x_2.

Bodyfat example

In `proc reg` you can get R^2_{Y1}, $R^2_{Y2|1}$, and $R^2_{Y3|12}$ by adding `pcorr1` as a `model` option. you can get $R^2_{Y1|23}$, $R^2_{Y2|13}$, and $R^2_{Y3|12}$ by adding `pcorr2`.

```
proc reg data=body;
 model bodyfat=triceps thigh midarm / pcorr1;
 model bodyfat=triceps thigh midarm / pcorr2; run;
-----------------------------------------------------------------------------------
                                Parameter Estimates

                                                                           Squared
                     Parameter       Standard                              Partial
Variable     DF       Estimate          Error    t Value    Pr > |t|   Corr Type I
Intercept     1      117.08469       99.78240       1.17      0.2578             .
triceps       1        4.33409        3.01551       1.44      0.1699       0.71110
thigh         1       -2.85685        2.58202      -1.11      0.2849       0.23176
midarm        1       -2.18606        1.59550      -1.37      0.1896       0.10501

                                                                            Squared
                     Parameter       Standard                               Partial
Variable     DF       Estimate          Error    t Value    Pr > |t|   Corr Type II
Intercept     1      117.08469       99.78240       1.17      0.2578              .
triceps       1        4.33409        3.01551       1.44      0.1699        0.11435
thigh         1       -2.85685        2.58202      -1.11      0.2849        0.07108
midarm        1       -2.18606        1.59550      -1.37      0.1896        0.10501
-----------------------------------------------------------------------------------
```

7.6 Multicollinearity

Recall: In the body fat example, the F-test for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ was *highly* significant, but individual t-tests for dropping any of x_1, x_2, or x_3 were *not* significant!

The set x_1, x_2, x_3 are useful for explaining body fat, but none of the three are useful *in the presence of the other two*.

Why? The predictors are measuring similar phenomena; their sample values are highly correlated. For example, $r = 0.924$ between triceps thickness x_1 and thigh circumference x_2.

This is known as *multicollinearity* among the predictors.

Effects of multicollinearity

- Model may still provide a good fit and precise prediction/estimation of the response.
- Several estimated regression coefficients $b_1, b_2, \ldots, b_k$ will have large standard errors, leading to conclusions that individual predictors are *not significant* although overall F-test may be *highly* significant.
- Concept of “holding all other predictors constant” doesn’t make sense in practice.
- Signs of regression coefficients may be “opposite” of intuition (or what we might think *marginally* they might be based on a scatterplot).

Bodyfat example

```
proc glm data=body;
 model bodyfat=triceps thigh midarm / solution; run;

                                               Sum of
Source                      DF          Squares     Mean Square    F Value     Pr > F
Model                        3      396.9846118     132.3282039      21.52     <.0001
Error                       16       98.4048882       6.1503055
Corrected Total             19      495.3895000

R-Square      Coeff Var      Root MSE     bodyfat Mean
0.801359       12.28017      2.479981         20.19500

                                         Standard
Parameter          Estimate                 Error    t Value    Pr > |t|
Intercept       117.0846948           99.78240295       1.17      0.2578
triceps           4.3340920            3.01551136       1.44      0.1699
thigh            -2.8568479            2.58201527      -1.11      0.2849
midarm           -2.1860603            1.59549900      -1.37      0.1896
```

Two of the three regression effects are *negative*. Holding midarm and triceps constant, increasing the thigh circumference 1 mm *decreases* bodyfat. Does this make sense?

Detecting multicollinearity

Predictor x_j has a *variance inflation factor* of

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the R^2 from regressing x_j on the remaining predictors $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$.
High R_j^2 (near 1) $\Rightarrow$ x_j is linearly associated with other predictors $\Rightarrow$ high VIF_j.

- $VIF_j \approx 1 \Rightarrow x_j$ is not involved in any multicollinearity.
- $VIF_j > 10 \Rightarrow x_j$ is involved in severe multicollinearity.

VIF$_j$'s in SAS

```
model bodyfat = triceps thigh midarm / vif;
-------------------------------------------------------------------------------------
                                Parameter Estimates

                        Parameter      Standard                           Variance
Variable      DF         Estimate         Error    t Value    Pr > |t|   Inflation
Intercept      1        117.08469      99.78240       1.17      0.2578           0
triceps        1          4.33409       3.01551       1.44      0.1699   708.84291
thigh          1         -2.85685       2.58202      -1.11      0.2849   564.34339
midarm         1         -2.18606       1.59550      -1.37      0.1896   104.60601
```

What do you conclude?

Remedies for multicollinearity

- Drop one or more predictors from the model. We'll discuss this in Chapter 9.
- More advanced: **principal components regression** uses indexes (new predictors) that are linear combinations of the original predictors as predictors in a new model. The indexes are selected to be uncorrelated. Disadvantage: the indexes might be hard to interpret.
- More advanced: **ridge regression** (Section 11.2).
- There is a handout on the course webpage giving more intuition behind the VIF_j if you are interested.